DARPA FEBRUARY 1992 ATIS BENCHMARK TEST RESULTS

David S. Pallett, Nancy L. Dahlgren, Jonathan G. Fiscus,
William M. Fisher, John S. Garofolo, Brett C. Tjaden

National Institute of Standards and Technology
Building 225, Room A216
Gaithersburg, MD 20899

1 INTRODUCTION

This paper documents the third in a series of Benchmark Tests for the DARPA Air Travel Information System (ATIS) common task domain. The first results in this series were reported at the June 1990 Speech and Natural Language Workshop [1], and the second at the February 1991 Speech and Natural Language Workshop [2]. The February 1992 Benchmark Tests include: (1) ATIS domain spontaneous speech recognition system tests, (2) ATIS natural language understanding tests, and (3) ATIS spoken language understanding tests.

Since the February 1991 tests, a large ATIS spoken language corpus has been collected, coordinated by a DARPA "Multi-Site ATIS Data COllection Working" (MADCOW) Group. The activities of this group, and NIST's role in that effort, are documented in another paper in this Proceedings [3].

2 OCTOBER 1991 "DRY RUN" TESTS

The procedures for test set selection, testing, scoring, adjudication, and reporting for the February 1992 ATIS Benchmark Tests were developed and used for a "dry run" test in October 1991, with unpublished results. A somewhat smaller test set was used at that time, which did not include test data from AT&T. The implementation of the tests was generally regarded as successful within the DARPA MADCOW Group and by the DARPA Spoken Language Program Coordinating Committee.

3 NEW CONDITIONS FOR THESE TESTS

The structure (and scoring) of these ATIS domain tests differ in several ways from the tests reported at the June 1990 and February 1991 Workshops:

- Following the February 1991 Workshop, minor revisions (e.g., to accommodate connecting flights, clarify terminology, revise headings and restructure tables, improve representation of fare structures, bug fixes, etc) were made to the relational air-travel-information database. The MADCOW data collection effort, and systems developed with this data, made use of this revised relational database (Version 3.3).
- The MADCOW data collection effort provided data from five sites (AT&T, BBN, CMU, MIT/LCS, and SRI), rather than the single ATIS data collection site (TI) used for the June 1990 and February 1991 tests.
- Some (but not all) of the collecting sites provided secondary (Crown PCC-160) microphone data in addition to the primary (close- talking Sennheiser) microphone. The use of the secondary microphone data was encouraged, but not required, for the February 1992 tests.
- The definition of "Class D" queries was broadened to include "Class D1" queries.
- The files indicating the "classification" (i.e., Class A, D or X) for each query were not provided along with the test queries (as they had been in previous tests), so that each site had no extra information regarding the context-dependency or answerability of each query.
- Similarly, "unanswerable" (Class X) queries were not identified when the test material was released. If system developers provided answers for these queries, they were not scored.
- No utterances were to be treated differently on the grounds of the presence of disfluencies such as false starts or restarts. In the February 1991 tests, these utterances were regarded as "Optional".
- Concern had been expressed at the February 1991 meeting that some sites might have chosen to "overgenerate" (by providing verbose) NL and SLS answers rather than provide more succinct answers. It was argued that "correct" answers should have at

least the information in the ".ref" files previously used in scoring answers, but no more than in some specified maximal answer. Bob Moore and Eric Jackson, at SRI, proposed and implemented an algorithmic procedure for deriving maximal reference answers (".rf2") from the NLParse-generated SQL files used to generate the .ref files. Bill Fisher at NIST subsequently modified the NIST comparator (used in scoring the NL and SLS results) to implement the new "minimum/maximum" scoring procedure. The Principles of Interpretation document was modified to accommodate these changes.

- Special reports were to be prepared by NIST to partition the tabulations of results according to the originating sites for the test data.
- Following completion of each phase of scoring the results, NIST was to prepare and make available to all participants both detailed and summary reports via anonymous ftp.
- Because there had been a recommendation to report results for all answerable queries in complete subject-scenarios (i.e., the material collected during one subject's working of one scenario), test material was to be provided to the testing sites in complete subject-scenarios. Emphasis was to be placed on analysis of the subset of "answerable" queries (i.e., Class A+D), rather than on the individual classes A and/or D. Further, the weighted error percentage (defined as twice the percentage of incorrect or "false" answers plus the percentage of "No_Answer" responses) was identified as preferable to the single-number "Score" reported at the February 1991 meeting (Score (%) = 100 (%) - [Weighted Error (%)])

4 TEST MATERIAL SELECTION AND DISTRIBUTION

With the approval of the MADCOW Group, NIST had reserved approximately 20% of the pooled MADCOW data for test purposes. NIST screened this data for the occurrence of truncated utterances, rejected the subject-scenarios that included these phenomena, and determined that there was a sufficient quantity of reserved potential test material to permit release of a test set consisting of approximately 200 utterances from each of the five MADCOW sites contributing data. NIST did not monitor the audio quality of the .wav files nor review the accuracy of the transcriptions, since no criteria for acceptability based on these have been defined, although in retrospect this might have simplified the adjudication process.

The test material, subsequent to deletion of some material during the adjudication process, consisted of 970 non-null (and 1 null) utterances in all classes. The number of distinct scenarios used by all subjects was 42, with a total of 37 subjects ("speakers") completing 122 subject-scenarios. There were 17 male subjects, and 20 were female. Seven of the 122 subject-scenarios used the "Common-1" scenario; however, the test material selected from BBN and CMU did not include any instances of this scenario. The average number of queries per subject-scenario was 8. The MIT subject-scenarios had an average number of 4.6 queries, and SRI and CMU each had an average number of 12.1 queries per subject-scenario. There were 508 lexemes represented in the test material. The average number of words per utterance was about 11.

After NIST selected the test material, it was produced on CD-ROM. The test disc (NIST Speech Disc T3-1.1) was distributed to the testing sites on Jan. 6, 1992.

Concurrent with preparation of the CD-ROMs, NIST staff and the "Annotation Group" at SRI initiated preparation of the annotation files required to implement scoring.

5 TEST PROCEDURE

Following completion of locally administered single-pass-per-system tests, participating sites submitted results for (at least) three ATIS tests: the SPeech RECognition (SPREC), Natural Language (NL) and Spoken Language System (SLS) tests.

The format for data submission via e-mail was specified by NIST and all "official" results were received at NIST by 6:00 AM on Jan. 20, 1992. As in previous ATIS tests, answer hypotheses were to be in the form of lexical SNOR (.lsn) files for the SPREC results and in Common Answer Specification (CAS) format files for the NL and SLS results. Each submission was to be accompanied by a text file for each system providing a system description following a suggested format.

6 TEST SCORING, ADJUDICATION AND REPORTING PROCEDURE

Upon receipt of the test results, NIST implemented preliminary scoring with a reference answer set including .cat, .ref and .rf2 files developed at NIST and SRI for the NL and SLS tests, and the "lexical SNOR" (.lsn) files derived from the detailed (.sro) transcriptions provided by the collecting sites for the SPREC tests. On Jan. 24, 1992, upon completion of the preliminary scoring and preparation of the required reports, NIST released the preliminary results by anonymous ftp.

A detailed and formal procedure was established at NIST at the MADCOW group's request for handling requests for adjudication.

The participating sites filed a total of 122 requests for adjudication, which were treated by NIST and the SRI Annotation group in a manner similar to that followed for the training data's bug reports. Some of these requests involved more than one utterance, or reported on more than one "bug" in an utterance, so that the number of unique utterances potentially affected by the requests for adjudication was 193, or approximately 19% of the test material.

Of these utterances, the adjudicators determined that 99 (51%) actually required one or more changes. "No Action" decisions were made for the remaining 49%.

NIST was advised by Francis Kubala at BBN during the adjudication period that some of the reference transcriptions used for scoring the SPREC test appeared to be inaccurate. NIST subsequently reviewed all of the transcriptions noted by Kubala and corrected them as deemed appropriate.

In addition to the 99 utterances noted as part of the formal requests for adjudication requiring changes to the annotations, 26 test utterances were identified by the adjudicators as requiring changes.

The final total of 125 utterances (12.9% of the entire test set) for which annotation changes were made includes the following breakdown (by category):

- 42 with software problems related to annotations or scoring (e.g., NLParse, batching, or Comparator bugs),
- 36 for which annotation errors had been made,
- 27 involved problems with the transcriptions developed at the originating sites, and
- 20 involved differences of opinion in applying the Principles of Interpretation or the use of context in interpreting the query.

Following completion of the adjudication process, NIST released a set of "Official" ATIS Benchmark Test results to the community on Feb. 5, 1992.

NIST was subsequently advised by Paramax that corrections to the reference answer set that were to have been made during the adjudication process did not appear to have been made. NIST and SRI determined that this had in fact been the case, and a total of 19 .rf2 files were corrected. The entire set of NL and SLS results were then re-scored, and a "Revised Official" set of results was made available to the community. Analysis of the differences between these two sets of "official" results shows that only 5 of Paramax's NL and 4 of their SLS answers were scored differently.

Paramax also noted, following release of the "Revised Official" results, that 20 of their NL as well as another 20 of their SLS answers were scored as "False" because of known limitations in the NIST official scoring software. NIST had determined that the degree to which Paramax's answers were affected by this known limitation was approximately ten times more severe than for any other site, and declined to alter the scoring software to accomodate Paramax's unusual responses. NIST encouraged Paramax to develop and document "unofficial" results [4] with slightly modified scoring software.

A "handout" was prepared for, and distributed at, the February 1992 Speech and Natural Language Workshop containing the System Descriptions provided by the participants and NIST's summaries of Benchmark Test results.

7 BENCHMARK TEST RESULTS AND DISCUSSION

7.1 ATIS SPeech RECognition (SPREC) Test Results:

7.1.1 Close-Talking Microphone

Table 1 presents a tabulation of the February 1992 ATIS spontaneous Speech RECognition (SPREC) test results.

Results are presented for a number of defined subsets of the utterances, with the utterance classes defined in the annotation process. The set Class A+D+X is the set of all utterances in all classes, consisting of 971 utterances. The set Class A+D includes all answerable utterances, 687 in all. Individual scores for the component subsets Class A, Class D, and Class X are also included. The utterances in Classes D and X tend to have a greater degree of disfluency than those in Class A. This factor may be reflected in the corresponding error rates, since the lowest subset error rates are to be found for Class A utterances, and the highest for Class X.

In the set of answerable queries, Class A+D, the word error ranges from 6.2% to 13.8%, and the "Utterance error rate" (corresponding approximately to "sentence error rate", but acknowledging the fact that some utterances consist of more than one sentence) range from 34.6% to 60.1%.

The lowest word error rate, in any of the subsets, 5.8%, is noted for the BBN system described in [5] for the subset of Class A utterances.

Table 2 presents a matrix tabulation of ATIS SPREC results for the set of answerable queries, Class A+D. This matrix form of tabulation of results was developed at the MADCOW group's request to shed light on potential variabilities in the data for test set components from differing originating sites. The five columns of the matrix block correspond to the five originating sites for the MADCOW test data. In this case, the six rows of the matrix block correspond to the six sets of SPREC test results sent to NIST. The "Overall Totals" column at the right of the central block presents results corresponding to those cited for the Class A+D subset in Table 1. Note, for example, that the previously cited lowest Class A+D subset word error of 6.2% (for the BBN system) is shown in the second row entry of this column.

The "Overall Totals" row presents results accumulated over all systems for which results were reported to NIST. Note that the Overall (subset) Total Word Error ("W. Error") ranges from a low of 5.9%, for the data originating at MIT/LCS, to 14.6% for the AT&T data subset.

These data suggest that the MIT data subset is less challenging for ATIS SPREC systems than the data from other sites, but the reasons for this are not immediately evident.

Analysis of the transcriptions suggests that the AT&T data subset has a higher incidence of disfluencies than other subsets, partially explaining why it is more challenging than the other data subsets.

For the "Class A+D" data, the lowest subset word error for any SPREC system is 3.2%, again for the BBN SPREC system and for the MIT data subset. Analysis of a similar matrix for the Class A data (not shown) indicates that the lowest subset word error (again for the MIT data subset) is 2.6% for the BBN system, with a corresponding utterance error of 20.7%.

7.1.2 Secondary (Crown PCC-160) Microphone Data

Three ATIS MADCOW sites provided data for both the Sennheiser close-talking microphone and the secondary (Crown PCC-160) microphone: CMU, MIT/LCS, and SRI. Two sites agreed to use the Crown microphone data with SPREC systems, using "robust" recognition algorithms: CMU and SRI. In some cases, results for other algorithms for comparable subsets of the data are available, and these have been excised from larger sets of data provided to NIST by CMU and SRI for the purposes of comparisons.

Table 3 presents a matrix tabulation of the SPREC data for the Class A+D data from CMU, MIT/LCS and SRI for 5 systems (i.e., 3 from CMU and 2 from SRI). The "cmu4" system is the CMU Sphinx II system [6] processing the close-talking microphone data, the "cmu6" system is the CMU codeword-dependent-cepstral-normalization (CDCN) system [7] processing the close-talking data, and the "cmu3" system is the CMU CDCN system processing the Crown microphone data. The "sri3" system (processing the close-talking microphone data) and "sri4" system (processing the Crown microphone data) are versions of the SRI Decipher system incorporating the "RASTA" procedure for high-pass filtering of a log-spectral representation of speech [8].

For the close-talking microphone data subset, the lowest word error rate (7.0%) is for the sri4 system, which may be compared to the cmu4 system (10.4%) and the cmu6 system (13.7%). According to the system description provided by CMU, the two CMU systems differ in the amount of training material, among other factors.

For the secondary microphone data subset, the word error rate for the cmu3 system is 17.8%, and for the sri4 system is 30.4%.

There are indications of substantial variabilities due to originating site for the secondary microphone data, with both the SRI and CMU data secondary microphone data subsets giving rise to higher error rates than for the MIT data subsets.

7.1.3 Statistical Significance: SPREC

As in previous benchmark tests, two statistical significance tests are routinely implemented at NIST in analysis of speech recognition performance assessment tests. The utterance (sentence) error test is an application of McNemar's test, first suggested for use in this community by Gillick [9]. Another test consists of a MAtched-Pairs Sentence-Segment Word Error ("MAPSSWE") significance test, originally devised for use with the Resource Management corpora.

Analysis of the tabulation of the word error test results for the answerable query subset (Class A+D) shown in Table 4a indicates that for the BBN system [5], the word error rates are significantly different from (lower than) those for the other systems included in these tests. The sentence error McNemar test (Table 4b) indicates a similar result, but in this case, the sentence error rate for the Paramax SPREC system [4] does not differ significantly from the BBN system.

7.2 Natural Language (NL) Tests

Table 5 presents a tabulation of the February 1992 ATIS Natural Language (NL) understanding tests results. Results are presented for the set of all "answerable utterances", Class A+D, and for the individual Class A and Class D subsets. As was the case for the SPREC results, in general the error rates are higher for Class D than for Class A utterances.

For the set of answerable queries, Class A+D, the weighted error ranges from 30.1% to 75.4%. Note that five of the systems have weighted error percentages between 30.1% and 33.9%.

Table 6 presents a matrix tabulation for the NL test results for the set of answerable queries, Class A+D. There were a total of 687 queries in this set. The numbers tabulated for this set in Table 5 appear in the "Overall Totals" column, along with corresponding percentages. The "Overall Totals" row indicates the variability due to the test subsets' originating site.

Of the 5 data subsets, the lower weighted error percentages in the "Overall Totals" row are to be found for the CMU and MIT data, with the SRI, AT&T, and BBN data giving rise to higher weighted error percentages.

Since the AT&T data was collected using a significantly different collection paradigm – with the subject interfacing with the ATIS system simulation only over a phone line, rather than viewing a screen display of travel information [10] – the fact that the AT&T data subset is more difficult than three other sites is perhaps not surprising.

However, the BBN ATIS data collection effort also differed somewhat from that at other MADCOW sites in that – although information was presented using a screen display – the BBN scenarios "included not only trip planning scenarios, but also problem solving involving more general kinds of database access... This was done to try to elicit a richer range of language usage.[3]" This factor ("richer language usage") may provide a partial explanation for the high NL error rates noted for the BBN data subset.

For the CMU and MIT [11] systems, there appears to be some indication that the error percentages for "locally-collected" data are lower than for "foreign" data, perhaps because of greater familiarity with the local data-collection scenarios and environment, or use of a variant of the system under test when collecting the MADCOW data from which the test set was selected.

7.3 Spoken Language Systems (SLS) Tests

Table 7 presents a tabulation of the February 1992 Spoken Language System understanding test results. As was the case for Table 5 (for the corresponding NL results), results are shown for several classes of the data, but emphasis in this material is placed on the answerable utterances, comprising Class A+D.

For the Class A+D set, the seven SLS systems have weighted error ranging from 43.7% to 90.2%. Note that four systems (from three sites: BBN, MIT and SRI [12]) have weighted error percentages between 43.7% and 52.8%.

Table 8 presents a matrix tabulation for the SLS test results for Class A+D, comparable in structure to that for the NL results of Table 6.

Of the 5 data subsets corresponding to different collection sites, the range in weighted error is from 49.5% (for the MIT data) to 73.1% (for the AT&T data).

8 ACKNOWLEDGEMENT

The authors would like to acknowledge the help provided by the entire MADCOW community throughout the ordeal of collecting, annotating, distributing and using the MADCOW corpus for test purposes. A companion paper in this Proceedings provides a detailed acknowledgement, but special credit was earned by Lynette Hirschman as Chair of the MADCOW Group. It is to everyone's credit that the essential data was collected, annotated, and distributed and that "deadlines" were usually honored!

Special thanks are also due to the group at MIT, particularly Michael Phillips and Christie Clark Winterton, for quick turn- around in producing recordable CD-ROM discs for distribution of the MADCOW training corpus from master tapes produced by Brett Tjaden at NIST.

The Annotation Group at SRI, consisting of Kate Hunicke-Smith, Harry Bratt and Beth Bryson, was invaluable to the NIST effort to implement these tests.

They participated actively and cheerfully in annotation of the test material and the adjudication process, in addition to "training" one of the authors (ND) in the use of the NLParse software and annotation techniques.

Francis Kubala, at BBN, called NIST's attention to some problematic transcriptions for the SPREC tests. NIST reviewed and revised these as appropriate, and in the process noted 3 truncated utterances (in one subject-scenario collected at BBN). While the revised transcriptions were used in NIST's "revised official" scoring, NIST neglected to delete this subject-scenario from the NL and SLS tests, as specified by MADCOW protocols for handling data with truncated utterances. Analysis of performance on this particular subject-scenario indicates that most sites did well, nonetheless.

9 References

1. Pallett, et al., "DARPA ATIS Test Results June 1990", in Proc. Speech and Natural Language Workshop, June 1990, (R. Stern, ed.) Morgan Kaufmann Publishers, Inc. ISBN 1-55860-157-0, pp. 114-121.
2. Pallett, D.S., "Session 2: DARPA Resource Management and ATIS Benchmark Test Poster Session", in Proc. Speech and Natural Language Workshop, February 1991, (P. Price, ed.) Morgan Kaufmann Publishers, Inc. ISBN 1-55860-207-0, pp. 49-58.
3. MADCOW, "Multi-Site Data Collection for a Spoken Language Corpus", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
4. Norton, L.M., Dahl, D.A., and Linebarger, M.C., "Recent Improvements and Benchmark Results for the Paramax ATIS System", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
5. Kubala, F. et al., "BBN BYBLOS and HARC February 1992 ATIS Benchmark Results", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
6. Ward, W. et al., "Speech Understanding in Open Tasks", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
7. Stern, R. M., et al., "Multiple Approaches to Robust speech Recognition" in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
8. Murveit, H., Butzberger, J. and Weintraub, M., "Reduced Channel Dependence for Speech Recognition", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
9. Gillick, L. and Cox, S.J., "Some Statistical Issues in the Comparison of speech Recognition Algorithms", Proceedings of ICASSP-89, Glasgow, May 1989, pp.532-535.
10. Pieraccini, R. et al., "Progress Report on the Chronus System: ATIS Benchmark Results", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
11. Zue, V., et al., "The MIT ATIS System: February 1992 Progress Report", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
12. Appelt, D.E. and Jackson, E., "SRI International February 1992 ATIS Benchmark Test Results", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.

10 APPENDIX: "OFFICIAL" VS. "UNOFFICIAL" RESULTS

Several sites expressed interest in having results for additional systems included in NIST's "official" summary, although these results typically were not available at the required time for "official" scoring. At least one site took exception to an idiosyncratic property of the "official" comparator's treatment of their system's responses to several queries, and requested permission to present "unofficial" results at the meeting. Another site noted that they had identified a "bug" in their CAS-answer- format software, and after it was fixed, they also requested permission to report unofficial results.

It was subsequently decided that the results submitted to NIST by the specified deadline, and uniformly scored at NIST with the "official" comparator and the adjudicated final set of reference answers would comprise the only "official" results, and that locally scored results should be represented as "unofficial", even if scored with the same scoring software and answer set as the "official" results.

It should be noted that since the results are for locally implemented tests, and since NIST's role in the tests is principally one of selecting and distributing the test material, and implementing the scoring software and uniformly tabulating the results of the tests, the results are not to be construed or represented as endorsements of any systems or official findings on the part of NIST, DARPA or the U.S. Government.

Class A+D+X Subset

	Corr	Sub	Del	Ins	Err	U. Err	# Utt.	
att3-adx	85.6	10.5	3.9	3.1	17.5	64.6	970	ATT Feb 92 Sprec Results
bbn3-adx	92.5	5.7	1.8	1.8	9.4	40.3	971	BBN Feb 92 Sprec Results
cmu4-adx	88.2	9.7	2.1	4.4	16.2	60.2	971	CMU Feb 92 ATIS Sphinx-II Senn.
mit4-adx	84.1	11.5	4.4	2.3	18.1	59.6	971	MIT-LCS Feb 92 Sprec Results
paramax3-adx	91.5	6.3	2.1	2.1	10.6	42.2	971	Paramax/BBN Feb 92 Sprec Results
sri3-adx	91.4	6.8	1.8	2.4	11.0	48.7	971	SRI Feb 92 Sprec Results

Class A+D Subset

	Corr	Sub	Del	Ins	Err	U. Err	# Utt.	
att3-a_d	88.9	7.7	3.4	2.7	13.8	60.1	687	ATT Feb 92 Sprec Results Class A+D
bbn3-a_d	95.2	3.6	1.1	1.5	6.2	34.6	687	BBN Feb 92 Sprec Results Class A+D
cmu4-a_d	91.9	6.5	1.6	3.7	11.8	54.4	687	CMU Feb 92 ATIS Sphinx-II Senn. Class A+D
mit4-a_d	88.3	8.7	3.1	1.9	13.6	54.1	687	MIT-LCS Feb 92 Sprec Results Class A+D
paramax3-a_d	94.6	4.0	1.4	1.7	7.1	36.4	687	Paramax/BBN Feb 92 Sprec Results Class A+D
sri3-a_d	93.8	4.9	1.4	2.1	8.4	44.5	687	SRI Feb 92 Sprec Results Class A+D

Class A Subset

	Corr	Sub	Del	Ins	Err	U. Err	# Utt.	
att3-a	88.9	7.2	3.9	2.0	13.1	60.9	402	ATT Feb 92 Sprec Results Class A
bbn3-a	95.4	3.3	1.3	1.2	5.8	35.6	402	BBN Feb 92 Sprec Results Class A
cmu4-a	92.8	5.7	1.6	3.2	10.4	54.2	402	CMU Feb 92 ATIS Sphinx-II Senn. Class A
mit4-a	89.1	7.8	3.1	1.6	12.5	54.5	402	MIT-LCS Feb 92 Sprec Results Class A
paramax3-a	94.9	3.6	1.5	1.4	6.5	36.6	402	Paramax/BBN Feb 92 Sprec Results Class A
sri3-a	94.4	4.0	1.5	1.7	7.3	44.0	402	SRI Feb 92 Sprec Results Class A

Class D Subset

	Corr	Sub	Del	Ins	Err	U. Err	# Utt.	
att3-d	89.0	8.7	2.3	4.1	15.2	58.9	285	ATT Feb 92 Sprec Results Class D
bbn3-d	94.9	4.2	0.8	1.9	7.0	33.3	285	BBN Feb 92 Sprec Results Class D
cmu4-d	90.3	8.2	1.5	4.8	14.5	54.7	285	CMU Feb 92 ATIS Sphinx-II Senn. Class D
mit4-d	86.7	10.3	3.0	2.3	15.7	53.7	285	MIT-LCS Feb 92 Sprec Results Class D
paramax3-d	94.1	4.7	1.1	2.2	8.1	36.1	285	Paramax/BBN Feb 92 Sprec Results Class D
sri3-d	92.5	6.4	1.1	2.8	10.3	45.3	285	SRI Feb 92 Sprec Results Class D

Class X Subset

	Corr	Sub	Del	Ins	Err	U. Err	# Utt.	
att3-x	77.4	17.3	5.3	3.9	26.5	75.6	283	ATT Feb 92 Sprec Results Class X
bbn3-x	85.5	11.0	3.5	2.7	17.2	53.9	284	BBN Feb 92 Sprec Results Class X
cmu4-x	78.9	17.6	3.4	6.1	27.2	74.3	284	CMU Feb 92 ATIS Sphinx-II Senn. Class X
mit4-x	73.8	18.5	7.7	3.3	29.5	72.9	284	MIT-LCS Feb 92 Sprec Results Class X
paramax3-x	83.7	12.2	4.0	3.1	19.4	56.3	284	Paramax/BBN Feb 92 Sprec Results Class X
sri3-x	85.5	11.5	3.0	2.9	17.4	58.8	284	SRI Feb 92 Sprec Results Class X

Table 1: ATIS SPREC Test Results

		Class A+D Subset Originating Site of Test Data					Overall Totals 687	Foreign Coll. Site Totals
		ATT (114 Utt.)	BBN (151 Utt.)	CMU (137 Utt.)	MIT (152 Utt.)	SRI (133 Utt.)		
SYSTEMS	att3	13.0 3.2 4.9 21.0 69.3	6.9 2.7 1.5 11.1 55.0	7.3 6.8 2.0 16.0 75.9	4.4 1.3 2.0 7.8 42.1	8.7 2.9 4.0 15.6 62.4	7.7 3.4 2.7 13.8 60.1	6.7 3.4 2.3 12.4 58.3
	bbn3	6.3 1.1 3.0 10.4 50.9	3.3 0.8 1.2 5.3 31.1	3.2 1.6 1.0 5.8 38.7	1.8 0.7 0.8 3.2 21.7	4.5 1.4 1.7 7.7 35.3	3.6 1.1 1.5 6.2 34.6	3.7 1.2 1.5 6.5 35.6
	cmu4	10.2 1.4 7.3 18.9 68.4	4.8 1.5 1.5 7.9 47.0	7.0 2.1 5.2 14.3 69.3	3.9 1.1 1.2 6.3 44.1	7.9 1.7 4.4 14.0 47.4	6.5 1.6 3.7 11.8 54.4	6.4 1.4 3.3 11.1 50.7
	mit4	7.9 2.5 3.1 13.4 51.8	7.6 3.6 1.5 12.7 57.6	9.9 3.9 1.7 15.5 61.3	5.9 1.8 0.9 8.6 46.1	13.1 3.7 2.6 19.4 54.1	8.7 3.1 1.9 13.6 54.1	9.5 3.5 2.2 15.1 56.4
	paramax3	6.8 1.5 2.8 11.0 48.2	3.1 0.7 1.2 5.0 28.5	3.4 2.3 1.2 6.9 41.6	2.5 0.7 1.1 4.3 26.3	5.2 1.9 2.6 9.7 41.4	4.0 1.4 1.7 7.1 36.4	4.0 1.4 1.7 7.1 36.4
	sri3	8.1 1.3 3.6 13.1 57.0	3.2 1.2 1.5 5.9 35.8	5.4 1.9 2.7 10.0 56.2	3.2 1.3 1.0 5.5 40.8	5.2 1.1 2.2 8.6 36.1	4.9 1.4 2.1 8.4 44.5	4.8 1.4 2.1 8.3 46.6
Overall Totals		8.7 1.8 4.1 14.6 57.6	4.8 1.8 1.4 8.0 42.5	6.0 3.1 2.3 11.4 57.2	3.6 1.1 1.2 5.9 36.8	7.4 2.1 2.9 12.5 46.1		
Foreign System		7.9 1.5 4.0 13.4 55.3	5.1 1.9 1.4 8.5 44.8	5.8 3.3 1.7 10.8 54.7	3.2 1.0 1.2 5.4 35.0	7.9 2.3 3.1 13.3 48.1		

%Sub %Del %Ins %W.Err %Utt.Err

Table 2: ATIS SPREC Results Class A+D by Collection Site

		Class A+D Subset Originating Site of Test Data			Overall Totals 332	Foreign Coll. Site Totals
		CMU (101 Utt.)	MIT (152 Utt.)	SRI (79 Utt.)		
SYSTEMS	cmu3	14.9 3.4 6.9 25.2 84.2	8.3 2.9 1.3 12.4 61.8	11.1 3.9 4.6 19.6 54.4	10.9 3.3 3.7 17.8 66.9	9.1 3.2 2.2 14.5 59.3
	cmu4	7.6 1.4 6.6 15.5 70.3	3.9 1.1 1.2 6.3 44.1	6.4 0.8 5.5 12.7 43.0	5.6 1.1 3.7 10.4 51.8	4.7 1.1 2.4 8.1 43.7
	cmu6	9.1 1.4 11.7 22.1 80.2	5.0 1.3 2.7 8.9 53.3	6.0 1.1 5.3 12.5 50.6	6.4 1.3 6.0 13.7 60.8	5.3 1.2 3.4 9.9 52.4
	sri3	5.2 1.5 2.7 9.4 52.5	3.2 1.3 1.0 5.5 40.8	3.8 1.1 2.2 7.1 34.2	4.0 1.3 1.8 7.0 42.8	4.0 1.3 1.6 7.0 45.5
	sri4	26.0 2.7 17.4 46.2 93.1	14.1 3.4 3.3 20.8 78.9	17.1 4.1 8.4 29.6 77.2	18.4 3.3 8.7 30.4 82.8	18.7 3.2 8.7 30.6 84.6
Overall Totals		12.6 2.1 9.1 23.7 76.0	6.9 2.0 1.9 10.8 55.8	8.9 2.2 5.2 16.3 51.9		
Foreign System		15.6 2.1 10.1 27.8 72.8	6.9 2.0 1.9 10.8 55.8	7.8 2.0 5.1 14.9 49.4		

%Sub %Del %Ins %W.Err %Utt.Err

Table 3: ATIS SPREC Crown and Crown Subset of Sennheiser Test Results by Collection sites

```
                  COMPARISON MATRIX: FOR THE MATCHED PAIRS TEST
                      Feb 91 ATIS SPREC Class A+D Results
                    Minimum Number of Correct Boundary words 2

|---------------------------------------------------------------------|
|          | att3-a_d | bbn3-a_d | cmu4-a_d | mit4-a_d | paramax3 | sri3-a_d |
|----------+----------+----------+----------+----------+----------+----------|
| att3-a_d |          | bbn3-a_d | cmu4-a_d |   same   | paramax3 | sri3-a_d |
|----------+----------+----------+----------+----------+----------+----------|
| bbn3-a_d |          |          | bbn3-a_d | bbn3-a_d | bbn3-a_d | bbn3-a_d |
|----------+----------+----------+----------+----------+----------+----------|
| cmu4-a_d |          |          |          | cmu4-a_d | paramax3 | sri3-a_d |
|----------+----------+----------+----------+----------+----------+----------|
| mit4-a_d |          |          |          |          | paramax3 | sri3-a_d |
|----------+----------+----------+----------+----------+----------+----------|
| paramax3 |          |          |          |          |          | paramax3 |
|----------+----------+----------+----------+----------+----------+----------|
| sri3-a_d |          |          |          |          |          |          |
|---------------------------------------------------------------------|
```

```
             COMPARISON MATRIX: McNEMAR'S TEST ON CORRECT SENTENCES FOR THE TEST:
                           Feb 91 ATIS SPREC Class A+D Results
                                     For all systems

|-------------------------------------------------------------------------------------------------|
|              | att3-a_d(274)| bbn3-a_d(449)| cmu4-a_d(313)| mit4-a_d(315)| paramax3(437)| sri3-a_d(381)|
|--------------+--------------+--------------+--------------+--------------+--------------+--------------|
| att3-a_d(274)|              |   D=(175)    |   D=( 39)    |   D=( 41)    |   D=(163)    |   D=(107)    |
|              |              |   bbn3-a_d   |   cmu4-a_d   |   mit4-a_d   |   paramax3   |   sri3-a_d   |
|--------------+--------------+--------------+--------------+--------------+--------------+--------------|
| bbn3-a_d(449)|              |              |   D=(136)    |   D=(134)    |   D=( 12)    |   D=( 68)    |
|              |              |              |   bbn3-a_d   |   bbn3-a_d   |     same     |   bbn3-a_d   |
|--------------+--------------+--------------+--------------+--------------+--------------+--------------|
| cmu4-a_d(313)|              |              |              |   D=(  2)    |   D=(124)    |   D=( 68)    |
|              |              |              |              |     same     |   paramax3   |   sri3-a_d   |
|--------------+--------------+--------------+--------------+--------------+--------------+--------------|
| mit4-a_d(315)|              |              |              |              |   D=(122)    |   D=( 66)    |
|              |              |              |              |              |   paramax3   |   sri3-a_d   |
|--------------+--------------+--------------+--------------+--------------+--------------+--------------|
| paramax3(437)|              |              |              |              |              |   D=( 56)    |
|              |              |              |              |              |              |   paramax3   |
|--------------+--------------+--------------+--------------+--------------+--------------+--------------|
| sri3-a_d(381)|              |              |              |              |              |              |
|              |              |              |              |              |              |              |
|-------------------------------------------------------------------------------------------------|
```

Table 4: ATIS SPREC Significance Test Comparisons: Class A+D

Class A+D

system	# T	# F	# NA	# Utt	W. Err	Description
att1	378	209	100	687	75.4	ATT Feb92 ATIS
bbn1	527	73	87	687	33.9	BBN Feb92 ATIS
cmu1	582	102	3	687	30.1	CMU-Phoenix Feb92 ATIS
cmu8	560	101	26	687	33.2	CMU-MINDS-II Feb92 ATIS
mit2	551	87	49	687	32.5	MIT Feb92 ATIS
paramax1	311	122	254	687	72.5	PARAMAX Feb92 ATIS
sri1	533	60	94	687	31.1	SRI Feb92 ATIS

Class A

system	# T	# F	# NA	# Utt	W. Err	Description
att1-a	256	96	50	402	60.2	ATT Feb92 ATIS Class A NL
bbn1-a	322	26	54	402	26.4	BBN Feb92 ATIS Class A NL
cmu1-a	356	46	0	402	22.9	CMU-Phoenix Feb92 ATIS Class A NL
cmu8-a	346	46	10	402	25.4	CMU-MINDS-II Feb92 ATIS Class A NL
mit2-a	342	34	26	402	23.4	MIT Feb92 ATIS Class A NL
paramax1-a	223	50	129	402	57.0	PARAMAX Feb92 ATIS Class A NL
sri1-a	335	25	42	402	22.9	SRI Feb92 ATIS Class A NL

Class D

system	# T	# F	# NA	# Utt	W. Err	Description
att1-d	122	113	50	285	96.8	ATT Feb92 ATIS Class D NL
bbn1-d	205	47	33	285	44.6	BBN Feb92 ATIS Class D NL
cmu1-d	226	56	3	285	40.4	CMU-Phoenix Feb92 ATIS Class D NL
cmu8-d	214	55	16	285	44.2	CMU-MINDS-II Feb92 ATIS Class D NL
mit2-d	209	53	23	285	45.3	MIT Feb92 ATIS Class D NL
paramax1-d	88	72	125	285	94.4	PARAMAX Feb92 ATIS Class D NL
sri1-d	198	35	52	285	42.8	SRI Feb92 ATIS Class D NL

Table 5: Feb 92 ATIS NL Test Results - Using Minimal/Maximal Scoring Criterion

```
-----------------------------------------------------------------------------------------------------------------
|            |                      Class (A+D) Set                                ||             |             |
|            |                 Originating Site of Test Data                       ||   Overall   |   Foreign   |
|            |     ATT     |     BBN     |     CMU     |     MIT     |     SRI     ||   Totals    | Coll. Site  |
|            |     114     |     151     |     137     |     152     |     133     ||     687     |   Totals    |
|------------+-------------+-------------+-------------+-------------+-------------||-------------+-------------|
|   att1     |  60  39  15 |  69  36  46 |  80  46  11 |  98  43  11 |  71  45  17 || 378 209 100 | 318 170  85 |
|            |  53  34  13 |  46  24  30 |  58  34   8 |  64  28   7 |  53  34  13 ||  55  30  15 |  55  30  15 |
|            |    81.6     |    78.1     |    75.2     |    63.8     |    80.5     ||    75.4     |    74.2     |
|   ---------+-------------+-------------+-------------+-------------+-------------||-------------+-------------|
|   bbn1     |  84  13  17 | 120  19  12 |  98  10  29 | 130  11  11 |  95  20  18 || 527  73  87 | 407  54  75 |
|            |  74  11  15 |  79  13   8 |  72   7  21 |  86   7   7 |  71  15  14 ||  77  11  13 |  76  10  14 |
|            |    37.7     |    33.1     |    35.8     |    21.7     |    43.6     ||    33.9     |    34.1     |
|   ---------+-------------+-------------+-------------+-------------+-------------||-------------+-------------|
|   cmu1     |  99  14   1 | 110  41   0 | 125  10   2 | 134  18   0 | 114  19   0 || 582 102   3 | 457  92   1 |
|            |  87  12   1 |  73  27   0 |  91   7   1 |  88  12   0 |  86  14   0 ||  85  15   0 |  83  17   0 |
| S          |    25.4     |    54.3     |    16.1     |    23.7     |    28.6     ||    30.1     |    33.6     |
| Y ---------+-------------+-------------+-------------+-------------+-------------||-------------+-------------|
| S cmu8     |  89  13  12 | 107  41   3 | 122  10   5 | 131  19   2 | 111  18   4 || 560 101  26 | 438  91  21 |
| T          |  78  11  11 |  71  27   2 |  89   7   4 |  86  12   1 |  83  14   3 ||  82  15   4 |  80  17   4 |
| E          |    33.3     |    56.3     |    18.2     |    26.3     |    30.1     ||    33.2     |    36.9     |
| M ---------+-------------+-------------+-------------+-------------+-------------||-------------+-------------|
| S mit2     |  83  19  12 | 114  26  11 | 111  14  12 | 137  10   5 | 106  18   9 || 551  87  49 | 414  77  44 |
|            |  73  17  11 |  75  17   7 |  81  10   9 |  90   7   3 |  80  14   7 ||  80  13   7 |  77  14   8 |
|            |    43.9     |    41.7     |    29.2     |    16.4     |    33.8     ||    32.5     |    37.0     |
|   ---------+-------------+-------------+-------------+-------------+-------------||-------------+-------------|
|   paramax1 |  36  22  56 |  58  27  66 |  89  18  30 |  71  26  55 |  57  29  47 || 311 122 254 | 311 122 254 |
|            |  32  19  49 |  38  18  44 |  65  13  22 |  47  17  36 |  43  22  35 ||  45  18  37 |  45  18  37 |
|            |    87.7     |    79.5     |    48.2     |    70.4     |    78.9     ||    72.5     |    72.5     |
|   ---------+-------------+-------------+-------------+-------------+-------------||-------------+-------------|
|   sri1     |  76  13  25 | 116  11  24 | 119  10   8 | 129  16   7 |  93  10  30 || 533  60  94 | 440  50  64 |
|            |  67  11  22 |  77   7  16 |  87   7   6 |  85  11   5 |  70   8  23 ||  78   9  14 |  79   9  12 |
|            |    44.7     |    30.5     |    20.4     |    25.7     |    37.6     ||    31.1     |    29.6     |
|================================================================================================================
| Overall    | 527 133 138 | 694 201 162 | 744 118  97 | 830 143  91 | 647 159 125 ||
| Totals     |  66  17  17 |  66  19  15 |  78  12  10 |  78  13   9 |  69  17  13 ||
|            |    50.6     |    53.4     |    34.7     |    35.4     |    47.6     ||         Legend:
|------------+-------------+-------------+-------------+-------------+-------------||   ------------------------
| Foreign    | 467  94 123 | 574 182 150 | 497  98  90 | 693 133  86 | 554 149  95 ||   | #T      #F   #NA     |
| System     |  68  14  18 |  63  20  17 |  73  14  13 |  76  15   9 |  69  19  12 ||   | %T      %F   %NA     |
| Totals     |    45.5     |    56.7     |    41.8     |    38.6     |    49.2     ||   | % Weighted Error     |
|----------------------------------------------------------------------------------    ------------------------
```

Table 6: ATIS NL Test Results - Using Minimal/Maximal Scoring Criterion

Class A+D

system	# T	# F	# NA	# Utt	W. Err	Description
att2	300	233	154	687	90.2	ATT Feb 92 ATIS SLS
bbn2	493	106	88	687	43.7	BBN Feb 92 ATIS SLS
cmu2	458	226	3	687	66.2	CMU Feb 92 ATIS SLS
mit1	471	132	84	687	50.7	MIT/SRI Feb 92 ATIS SLS
mit3	419	95	173	687	52.8	MIT Feb 92 ATIS SLS
paramax2	302	148	237	687	77.6	PARAMAX/BBN Feb 92 ATIS SLS
sri2	444	69	174	687	45.4	SRI Feb 92 ATIS SLS

Class A

system	# T	# F	# NA	# Utt	W. Err	Description
att2-a	208	118	76	402	77.6	ATT Feb 92 ATIS SLS Class A
bbn2-a	301	43	58	402	35.8	BBN Feb 92 ATIS SLS Class A
cmu2-a	298	104	0	402	51.7	CMU Feb 92 ATIS SLS Class A
mit1-a	305	58	39	402	38.6	MIT/SRI Feb 92 ATIS SLS Class A
mit3-a	288	47	67	402	40.0	MIT Feb 92 ATIS SLS Class A
paramax2-a	215	70	117	402	63.9	PARAMAX/BBN Feb 92 ATIS SLS Class A
sri2-a	305	32	65	402	32.1	SRI Feb 92 ATIS SLS Class A

Class D

system	# T	# F	# NA	# Utt	W. Err	Description
att2-d	92	115	78	285	108.1	ATT Feb 92 ATIS SLS Class D
bbn2-d	192	63	30	285	54.7	BBN Feb 92 ATIS SLS Class D
cmu2-d	160	122	3	285	86.7	CMU Feb 92 ATIS SLS Class D
mit1-d	166	74	45	285	67.7	MIT/SRI Feb 92 ATIS SLS Class D
mit3-d	131	48	106	285	70.9	MIT Feb 92 ATIS SLS Class D
paramax2-d	87	78	120	285	96.8	PARAMAX/BBN Feb 92 ATIS SLS Class D
sri2-d	139	37	109	285	64.2	SRI Feb 92 ATIS SLS Class D

Table 7: ATIS SLS Test Results - Using Minimal/Maximal Scoring Criterion

		Class (A+D) Set Originating Site of Test Data					Overall Totals 687	Foreign Coll. Site Totals
		ATT 114	BBN 151	CMU 137	MIT 152	SRI 133		
	att2	36 50 28 32 44 25 112.3	61 41 49 40 27 32 86.8	69 42 26 50 31 19 80.3	84 48 20 55 32 13 76.3	50 52 31 38 39 23 101.5	300 233 154 44 34 22 90.2	264 183 126 46 32 22 85.9
	bbn2	72 23 19 63 20 17 57.0	113 21 17 75 14 11 39.1	95 13 29 69 9 21 40.1	122 18 12 80 12 8 31.6	91 31 11 68 23 8 54.9	493 106 88 72 15 13 43.7	380 85 71 71 16 13 45.0
S Y S T E M S	cmu2	73 38 3 64 33 3 69.3	82 69 0 54 46 0 91.4	98 39 0 72 28 0 56.9	113 39 0 74 26 0 51.3	92 41 0 69 31 0 61.7	458 226 3 67 33 0 66.2	360 187 3 65 34 1 68.5
	mit1	72 28 14 63 25 12 61.4	103 30 18 68 20 12 51.7	94 19 24 69 14 18 45.3	121 22 9 80 14 6 34.9	81 33 19 61 25 14 63.9	471 132 84 69 19 12 50.7	350 110 75 65 21 14 55.1
	mit3	68 21 25 60 18 22 58.8	81 15 55 54 10 36 56.3	88 19 30 64 14 22 49.6	110 14 28 72 9 18 36.8	72 26 35 54 20 26 65.4	419 95 173 61 14 25 52.8	309 81 145 58 15 27 57.4
	paramax2	36 23 55 32 20 48 88.6	52 33 66 34 22 44 87.4	87 27 23 64 20 17 56.2	74 30 48 49 20 32 71.1	53 35 45 40 26 34 86.5	302 148 237 44 22 34 77.6	302 148 237 44 22 34 77.6
	sri2	55 12 47 48 11 41 62.3	101 13 37 67 9 25 41.7	93 13 31 68 9 23 41.6	112 20 20 74 13 13 39.5	83 11 39 62 8 29 45.9	444 69 174 65 10 25 45.4	361 58 135 65 10 24 45.3
Overall Totals		412 195 191 52 24 24 72.8	593 222 242 56 21 23 64.9	624 172 163 65 18 17 52.9	736 191 137 69 18 13 48.8	522 229 180 56 25 19 68.5		
Foreign System Totals		376 145 163 55 21 24 66.2	480 201 225 53 22 25 69.2	526 133 163 64 16 20 52.2	505 155 100 66 20 13 53.9	439 218 141 55 27 18 72.3		

```
       Legend:
--------------------
| #T     #F   #NA  |
| %T     %F   %NA  |
| % Weighted Error |
--------------------
```

Table 8: ATIS SLS Test Results - Using Minimal/Maximal Scoring Criterion

Experiments in Evaluating Interactive Spoken Language Systems[1]

Joseph Polifroni, Lynette Hirschman, Stephanie Seneff, and Victor Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

As the DARPA spoken language community moves towards developing useful systems for interactive problem solving, we must explore alternative evaluation procedures that measure whether these systems aid people in solving problems within the task domain. In this paper, we describe several experiments exploring new evaluation procedures. To look at end-to-end evaluation, we modified our data collection procedure slightly in order to experiment with several objective task completion measures. We found that the task completion time is well correlated with the number of queries used. We also explored log file evaluation, where evaluators were asked to judge the clarity of the query and the correctness of the response based on examination of the log file. Our results show that seven evaluators were unanimous on more than 80% of the queries, and that at least 6 out of 7 evaluators agreed over 90% of the time. Finally, we applied these new procedures to compare two systems, one system requiring a complete parse and the other using the more flexible robust parsing mechanism. We found that these metrics could distinguish between these systems: there were significant differences in ability to complete the task, number of queries required to complete the task, and score (as computed through a log file evaluation) between the robust and the non-robust modes.

INTRODUCTION

For the first two years of the DARPA Spoken Language Program, common evaluation in the ATIS domain has been performed solely with the Common Answer Specification (CAS) protocol [4], whereby a system's performance is determined by comparing its output, expressed as a set of database tuples, with one or more predetermined reference answers [1]. The CAS protocol has the advantage that system evaluation can be carried out automatically, once the principles for generating the reference answers have been established and a corpus has been annotated accordingly. Since direct comparison across systems can be performed relatively easily with this procedure, we have been able to achieve cross fertilization of research ideas, leading to rapid research progress.

QUERY 1:	PLEASE LIST THE FLIGHT FROM PITTSBURGH TO BALTIMORE THAT WILL BE MADE BY SIX SEAT AIRPLANE ON JUNE TWENTIETH
RESPONSE 1:	There are no flights from pittsburgh to Baltimore leaving before 6:00 a.m. on Saturday June 20.
QUERY 2:	LIST THE NUMBER OF FIRST CLASS FLIGHTS AVAILABLE ON DELTA AIRLINES
RESPONSE 2:	279

Table 1: Two examples of queries from the February '92 Evaluation for which the CAS evaluation gives a misleading assessment of the system's ability to understand.

However, CAS evaluation is not without its share of shortcomings. One unfortunate outcome of evaluating systems based on *answers* retrieved from the flight database is that we cannot separately assess the system's understanding of the query from its ability to retrieve the correct information and to conform to the prescribed output format. In the best case, researchers may end up spending a considerable amount of time worrying about the form and substance of the answer, which has little to do with language understanding. In the worst case, the results of the evaluation may be down right misleading. Consider the two examples from the February 1992 test-set shown in Figure 1. For Query 1, the system misunderstood the phrase "by six" as meaning "before 6:00 a.m." Nonetheless, the answer is judged correct, because both the hypothesized and reference answers are the NULL set, i.e., no flights satisfy the set of constraints. For Query 2, the system found 279 flights, but the correct answer is 278. The erroneous extra flight is the one connecting flight in the database shared by two airlines, Delta and USAIR.

Another shortcoming of the present evaluation procedure is that it has no place for interactive dialogue. In a realistic application, the user and the computer are often partners in problem solving, in which the final solution may be best obtained by allowing both sides to take the initiative in the conversation. Since the hu-

[1] This research was supported by DARPA under Contract N00014-89-J-1332, monitored through the Office of Naval Research.

man/computer dialogue can vary widely from system to system, it is impossible to use the data collected from one system to evaluate another system without making available the computer's half of the conversation. Even then, the system being tested becomes an observer analyzing two sides of a conversation rather than a participant.

To be sure, the current evaluation protocol has served the community well. The refinements made during the last year have significantly improved its ability to provide an objective benchmark. However, as we continue to press forward in developing *useful* spoken language systems that can help us solve problems, we must correspondingly expand the battery of evaluation protocols to measure the effectiveness of these systems in accomplishing specific tasks.

At the March 1991 meeting of the SLS Coordinating Committee, a working group was formed with the specific goal of exploring methodologies that will help us evaluate if, and how well, a spoken language system accomplishes its task in the ATIS domain. The consensus of the working group was that, while we may not have a clear idea about how to evaluate overall system performance, it is appropriate to conduct experiments in order to gain experience. The purpose of this paper is to describe three experiments conducted at MIT over the past few months related to this issue. These experiments explored a number of objective and subjective evaluation metrics, and found some of them to be potentially helpful in determining overall system performance and usefulness.

END-TO-END EVALUATION

In order to carry out end-to-end evaluation, i.e., evaluation of overall task completion effectiveness, we must be able to determine precisely the task being solved, the correct answer(s), and when the subject is done. Once these factors have been specified, we can then compute some candidate measures and see if any of them are appropriate for characterizing end-to-end system performance.

While true measures of system performance will require a (near) real-time spoken language system, we felt that some preliminary experiments could be conducted within the context of our ATIS data collection effort [3,2]. In our data collection paradigm, a typist types in the subject's queries verbatim, after removing disfluencies. All subsequent processing is done automatically by the system. To collect data for end-to-end evaluation, we modified our standard data collection procedure slightly, by adding a specific scenario which has a unique answer. For this scenario, the subjects were asked to report the answer explicitly.

As a preliminary experiment, we used two simple scenarios. In one of them, subjects were asked to determine the type of aircraft used on a flight from Philadelphia to Denver that makes a stop in Atlanta and serves breakfast. Subjects were asked to end the scenario by saying "End scenario. The answer is" followed by a statement of the answer, e.g., "End scenario. The answer is Boeing 727." From the log files associated with the session scenario, we computed a number of objective measures, including the success of task completion, task completion time, the number of successful and the number of unsuccessful queries (producing a "no answer" message)[2].

Measurements	Mean	Std. Dev.
Total # of Queries Used	4.8	1.6
# of Queries with Error Messages	1.0	1.4
Time to Completion (s.)	166.1	66.0

Table 2: Objective end-to-end measures.

We collected data from 29 subjects and analyzed the data from 24 subjects[3]. All subjects were able to complete the task, and statistics on some of the objective measures are shown in Table 2.

Figure 1 displays scatter plots of the number of queries used by each subject as a function of the task completion time. A least-square fit of the data is superimposed. The number of queries used is well correlated with the task completion time ($R = 0.84$), suggesting that this measure may be appropriate for quantifying the usefulness of systems, at least within the context of our experiment. Also plotted are the number of queries that generated a "no answer" message. The correlation of this measure with task completion time is not as good ($R = 0.66$), possibly due to subjects' different problem solving strategies and abilities.

LOG FILE EVALUATION

We also conducted a different set of experiments to explore subject-based evaluation metrics. Specifically, we extracted from the log files pairs of subject queries and system responses in sequence, and asked evaluators to judge the clarity of the query (i.e., clear, unclear, or unintelligible) and the correctness of the response (correct, partially correct, incorrect, or "system generated an error message"). A program was written to enable evaluators to enter their answers on-line, and the results were tabulated automatically. We used seven evaluators for this experiment, all people from within our group. Four people had detailed knowledge of the system and the desig-

[2] The system generates a range of diagnostic messages, reporting that it cannot parse, or that it cannot formulate a retrieval query, etc.

[3] Data from the remaining subjects were not analyzed, since they have been designated by NIST as test material.

Scenario Number	System	% of Scenarios w/Solution	Solution Correct	Completion Time(s)	Number of Queries	% of Queries Correct	% of Queries Incorrect	% of Queries No Answer	DARPA Score
1	Robust	100	100	215	4.4	94	0	6	94
1	Full	86	71	215	4.7	70	0	30	70
2	Robust	100	88	478	8.6	66	25	8	41
2	Full	86	86	483	10.6	39	4	56	35
3	Robust	100	100	199	4.4	82	15	3	68
3	Full	88	88	376	8.0	42	0	58	42
4	Robust	100	71	719	11.7	71	22	6	49
4	Full	75	38	643	9.8	51	0	49	51
All	Robust	100	90	399	7.2	75	18	6	57
All	Full	83	70	434	8.3	48	1	51	47

Table 4: Mean metrics for robust and full parse systems, shown by scenario

in these experiments.

The next column of the same table shows the average number of queries for each scenario. Since these numbers appear to be well correlated with task completion time, they suffer from some of the same deficiencies.

Log File Score In order to measure the number of queries correctly answered by the system, two system developers independently examined each query/answer pair and judged the answer as correct, partially correct, incorrect, or unanswered, based on the evaluation program developed for the logfile evaluation. The system developers were in complete agreement 92% of the time. The cases of disagreement were examined to reach a compromise rating. This provided a quick and reasonably accurate way to assess whether the subjects received the information they asked for. The percentages of queries correctly answered, incorrectly answered, and unanswered, and the resulting DARPA score (i.e., % correct - % incorrect) are shown in the last four columns of Table 4.

Although not shown in Table 4, the overall ratio of correctly answered queries to those producing no answer was an order of magnitude higher for the robust parser (148:13) than for the non-robust parser (118:125). This was associated with an order-of-magnitude increase in the number of *incorrect* answers: 32 vs. 3 for the non-robust parser. However, the percentage of "no answer" queries seemed to be more critical in determining whether a subject succeeded with a scenario than the percentage of incorrect queries.

Debriefing Questionnaire Each subject received a debriefing questionnaire, which included a question asking for a comparison of the two systems used. Unfortunately, data were not obtained from the first five subjects. Of the ten subjects that responded, five preferred the robust system, one preferred the non-robust system, and the remaining ones expressed no preference.

Difficulty of Scenarios There was considerable variability among the scenarios in terms of difficulty. Scenario 4 turned out to be by far the most difficult one to solve, with only a little over half of the sessions being successfully completed[4]. Subjects were asked to "choose a date within the next week" and to be sure that the restrictions on their fare were acceptable. We intentionally did not expand the system to understand the phrase "within the next week" to mean "no seven-day advance purchase requirement," but instead required the user to determine that information through some other means. Also in Scenario 4, there were no available first class fares that would exactly cover two coach class fares. Scenarios 2 and 4 were intended to be more difficult than 1 and 3, and indeed they collectively had a substantially lower percentage of correct query answers than the other two scenarios, reflecting the fact that subjects were groping for ways to ask for information that the system would be able to interpret.

There was a wide variation across subjects in their ability to solve a given scenario, and in fact, subjects deviated substantially from our expectations. Several subjects did not read the instructions carefully and ignored or misinterpreted key restrictions in the scenario. For instance, one subject thought the "within the next week" requirement in Scenario 4 meant that he should *return* within a week of his departure. Some subjects had a weak knowledge of air travel; one subject assumed that the return trip would be on the same flight as the forward leg, an assumption which caused considerable confusion for the system.

The full parser and robust parser showed different strengths and weaknesses in specific scenarios. For example, in Scenario 3, the full parser often could not parse the expression "Boeing 757", but the robust parser had no trouble. This accounts in part for the large "win" of the robust parser in this scenario. Conversely, in Scenario 4, the robust parser misinterpreted expressions of the type "about two hundred dollars", treating "about two" as a time expression. This led the conversation badly astray in these cases, and perhaps accounts for the

[4] The other three scenarios were solved successfully on average nearly 90% of the time.